Entropy-Gradient Grounding: Training-Free Evidence Retrieval in Vision-Language Models

Marcel Gröpl*,[1], Jaewoo Jung*,[3], Seungryong Kim[3], Marc Pollefeys[1], and Sunghwan Hong[1,2]

[1]ETH Zürich [2]ETH AI Center [3]KAIST AI
https://entropy-gradient-grounding.github.io

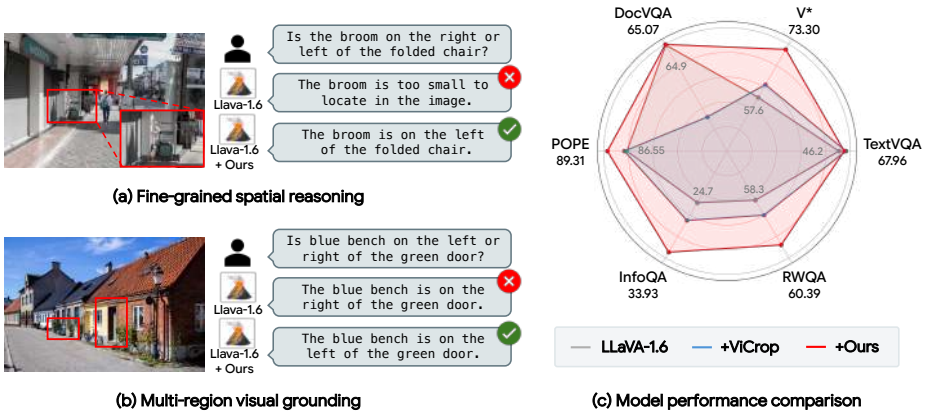

Fig. 1: Teaser. As shown in (a) and (b), existing VLMs struggle to answer questions when visual evidence is fine-grained or exists in spatially disjoint regions. We propose a training-free method where we apply a query-based visual grounding method to discover relevant regions and provide these regions as additional image crops, improving performance in both challenging scenarios.

Abstract. Despite rapid progress, pretrained vision–language models still struggle when answers depend on tiny visual details or on combining clues spread across multiple regions, as in documents and compositional queries. We address this by framing grounding as test-time evidence retrieval: given a query, the model should actively identify where to look next to resolve ambiguity. To this end, we propose a training-free, model-intrinsic grounding method that uses uncertainty as supervision. Specifically, we compute the entropy of the model's next-token distribution and backpropagate it to the visual token embeddings to obtain an *entropy-gradient* relevance map, without auxiliary detectors or attention-map heuristics. We then extract and rank multiple coherent regions to support multi-evidence queries, and introduce an iterative zoom-and-reground procedure with a spatial-entropy stopping rule to avoid over-refinement. Experiments on seven benchmarks across four VLM architectures demonstrate consistent improvements over existing methods, with the largest gains on detail-critical and high-resolution settings, while also producing more interpretable evidence localizations.

1 Introduction

Vision–Language Models (VLMs) have shown impressive performance on a broad range of multimodal tasks, including visual question answering and document understanding [4, 6, 7, 21, 31, 36]. Yet many failures persist when the decisive evidence is fine-grained, *e.g.,* small text, symbols, or scattered across disjoint regions, *e.g.,* documents, compositional queries [13, 37], as exemplified in Fig. 1. These failures suggest that the key challenge is not generation, but *selective perception*: identifying the right visual evidence. Accordingly, we cast such vision–language tasks as *active perception* at inference time: conditioned on a query, the model should perform grounding or decide *where to look* to acquire informative evidence.

A growing body of work explores *training-free* ways to extract grounding signals from pretrained VLMs [15, 37], avoiding additional training that can be prohibitively costly for today's large-scale models. Typically, many methods rely on attention maps for numerous downstream tasks [1, 2, 8, 9], but converting attention into reliable localization typically requires heuristic choices, *e.g.,* selecting heads/layers and designing post-processing, and the resulting maps can be brittle across backbones and dominated by a single salient region—a poor fit for queries that require aggregating spatially disjoint evidence. A complementary line of work *acts* on these imperfect signals by cropping or zooming into the predicted evidence at test time. ViCrop [37] exemplifies this direction, showing that zooming can improve downstream accuracy, but its region selection remains heuristic and it typically commits to a *single* crop. As a result, current approaches provide limited support for principled multi-region retrieval and controlled refinement: if the initial saliency is coarse or biased toward one cue, a one-shot zoom may miss secondary evidence, whereas exhaustive grid-based zooming can recover it only at substantial computational cost.

In this work, we propose a training-free, model-intrinsic grounding approach that treats inference as *test-time evidence acquisition* guided by the model's own uncertainty. Our key insight is that the entropy of the next-token distribution provides a decision-relevant indicator of visual evidence. Given an image and a query, we compute the entropy of the next-token distribution at a single decoding step and backpropagate it to the visual embeddings. The resulting *entropy-gradient* map highlights regions whose visual features most affect the model's uncertainty, producing a grounding signal that is tightly coupled to the model's prediction behavior—without auxiliary detectors or attention heuristics.

Crucially, our uncertainty-driven signal supports *multi-region* localization. We convert the entropy-gradient response into multiple spatially coherent regions of interest, enabling explicit aggregation over spatially disjoint evidence. To further recover fine-grained or secondary cues that can be suppressed in a single-pass map by dominant activations, we introduce an iterative refinement procedure that reapplies the same entropy-backpropagation mechanism on selected subregions. We regulate refinement with a spatial-entropy stopping criterion, terminating when spatial concentration no longer improves; empirically,

probability-based confidence measures are inconsistent for this control, while spatial entropy provides a stable convergence signal across backbones and tasks.

We evaluate our approach on multiple benchmarks [12, 19, 23, 24, 29, 32, 34] using strong pretrained backbones [5, 20, 21, 31]. Our method consistently improves over these baselines and outperforms the strongest training-free competitor, with gains that are particularly pronounced on detail-critical and multi-evidence queries. Qualitative results further show that our refinement yields more precise and interpretable evidence localization.

Our contributions are summarized as follows:

- We introduce a training-free, model-intrinsic grounding method for pretrained VLMs that treats inference as test-time evidence acquisition, using the model's own uncertainty to guide where to look.
- We develop a multi-region selection and ranking pipeline that extracts multiple coherent ROIs from entropy-gradient maps, enabling retrieval and aggregation of spatially disjoint evidence required by documents and compositional queries.
- We introduce an iterative refinement procedure that repeatedly re-grounds and re-crops selected regions, together with a spatial-entropy stopping criterion that provides early stopping.
- We provide extensive quantitative and qualitative evaluations across multiple benchmarks and demonstrate consistent improvements over baselines.

2 Related Work

Visual Grounding in MLLMs. Visual grounding aims to localize image regions relevant to a textual query. Classical approaches rely on supervised detectors, *e.g.,* YOLO [25], or open-vocabulary detectors that condition directly on text, *e.g.,* Grounding DINO [22]. Only limited work studies *training-free* grounding directly from pretrained MLLMs. ViCrop [37] extracts localization cues from attention-derived signals across forward passes, while other analyses probe attention heads and use heuristic criteria, *e.g.,* spatial entropy, to select informative heads [15]. Several methods instead introduce extra components to convert internal signals into masks, such as F-LMM [33] and LISA [17]. More generally, gradient-based attribution methods, *e.g.,* Grad-CAM [27], Integrated Gradients [30], indicate that prediction gradients encode spatial saliency, but they are primarily designed for post-hoc interpretability. In contrast, we derive grounding from gradients of an entropy-based objective over the language model's output distribution, yielding a training-free, model-intrinsic signal that enables principled *multi-region* localization without architectural modification.

Resolution Constraints and Region-Level Reasoning. Reasoning over fine-grained visual evidence remains challenging for MLLMs due to limited spatial granularity and token budgets. Many systems encode the image with fixed-resolution pipelines, e.g., LLaVA-1.5 [20], while others employ multi-crop or tiling strategies to increase visual coverage, e.g., LLaVA-1.6 [21], InternVL [31],

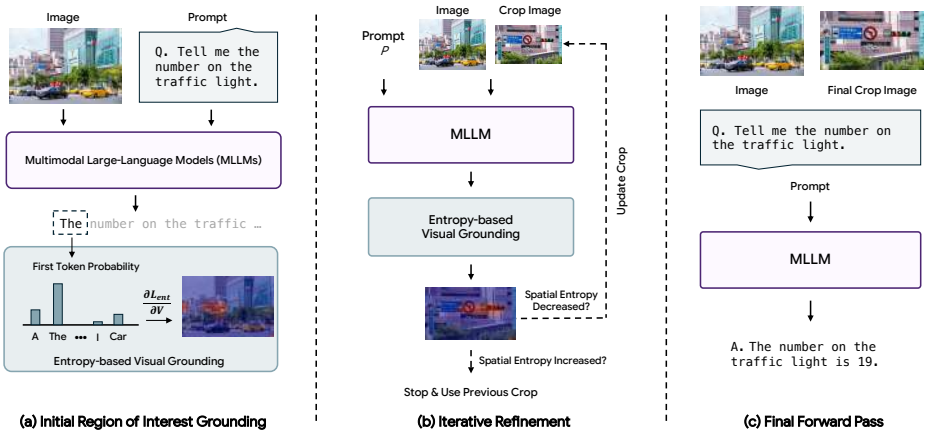

Fig. 2: Overview of proposed method. (a) Given an image and prompt, we obtain an initial region-of-interest by backpropagating the entropy of the next-token distribution to visual embeddings, producing an entropy-gradient relevance map. (b) We iteratively re-ground and re-crop the most informative regions, stopping when the spatial entropy criterion indicates further refinement no longer improves spatial concentration. (c) The final crop(s) are used for a forward pass to produce the answer.

Molmo [6]. Qwen-VL processes images at native resolution without predefined cropping [4], and LLaVA-OneVision [3] explores native-resolution encoding via RiCe-ViT [18, 35]. Beyond static resolution strategies, several works introduce explicit region-level reasoning. TEVA [13] predicts regions of interest using auxiliary detectors and additional training stages to guide patch selection and visual encoding. SEAL [32] performs LLM-guided visual search with dedicated control components and maintains a visual working memory. We also equip region-level reasoning to existing MLLMs by providing additional image crops discovered through our proposed training-free entropy-based visual grounding method.

3 Method

3.1 Problem Formulation and Overview

We consider a pretrained Multimodal Large Language Model (MLLM) that answers a query given an image–prompt pair (I, P). Our goal is to improve visual understanding via *training-free, model-intrinsic* grounding: for each (I, P), we seek to localize the spatial regions in I that constitute the most decision-relevant evidence for answering the query.

In the following, we first describe our entropy-based visual grounding method, which backpropagates the entropy of the next-token distribution to obtain an entropy-gradient map (Sec. 3.2). Since a single saliency map often collapses to a dominant region and fails to capture spatially disjoint evidence, we then introduce a multi-region selection and ranking procedure that extracts multiple coherent regions of interest from the gradient map (Sec. 3.3). Finally, to recover

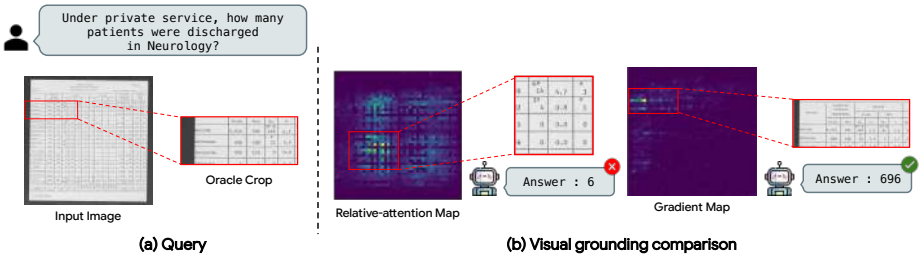

Fig. 3: Relative attention vs. entropy-gradient. Given an image and a user query as shown in (a), we compare the attention-based grounding used in ViCrop [37] with our proposed entropy-gradient grounding in (b). The relative-attention map produces a diffuse heatmap that highlights an incorrect region, leading to a wrong crop and an erroneous answer (6). In contrast, the entropy-gradient map concentrates on the query-relevant row of the table, yielding a precise crop that enables the model to correctly answer the question (696).

fine-grained or secondary cues and to provide a controlled feedback loop, we present an iterative refinement scheme regulated by a spatial-entropy stopping criterion (Sec. 3.4). An overview of our method is illustrated in Fig. 2.

3.2 Entropy-Based Visual Grounding

Our goal is to obtain a model-intrinsic grounding signal that reflects decision-relevant evidence for a given query. Prior works have explored text-to-image token attention maps to explain which visual regions the model attends to when generating a response [14–16, 37, 38]. However, extracting a clear, interpretable attention map typically requires heuristic selection of specific heads or layers, and often additional post-processing steps [15, 37]. Even then, as exemplified in Fig. 3-(b), the resulting maps may be inaccurate, reflect internal token routing rather than the visual evidence that actually drives the model's output, and often the selected attention layers and heads are model-specific, requiring independent analysis for each model.

Entropy as a grounding objective. Rather than aggregating signals spread across many attention operations, we derive grounding directly from the model's final prediction: its *next-token distribution*. We do so via gradient-based attribution [27]: we define a scalar objective on the output distribution and measure its sensitivity to the visual embeddings. Concretely, we use the Shannon entropy of the next-token distribution and backpropagate it to the visual embeddings, producing an *entropy-gradient* map that is model-intrinsic, query-conditioned, and training-free—requiring no head selection, auxiliary modules, or post-processing heuristics.

Formally, given an image I and prompt P, at decoding step t the model produces a next-token distribution $p_t(y) = p(y_t = y \mid I, P, y_{<t})$ over the vocabulary

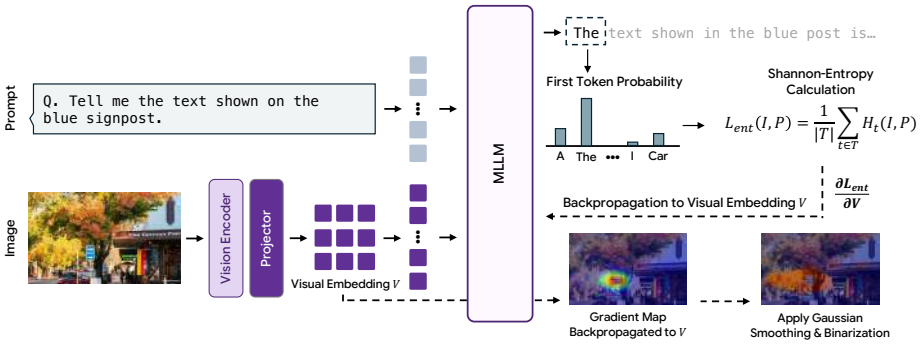

Fig. 4: Visual grounding via entropy-gradient map. Given an image and a text prompt, the image is encoded by a vision encoder and projected into visual embeddings V, which are fed alongside the prompt tokens into the MLLM. We compute the Shannon entropy from the next-token probability of the first token. This entropy objective is then backpropagated to the visual embeddings V. The resulting gradient map highlights image regions whose visual features most influence the model's predictive uncertainty. A Gaussian smoothing step followed by adaptive binarization converts the continuous map into a clean spatial mask, from which coherent regions of interest are extracted for downstream cropping and answer generation.

$\mathcal{Y}$. We define its Shannon entropy as

$$H_t(I,P) \;=\; -\sum_{y\in\mathcal{Y}} p_t(y)\,\log p_t(y). \tag{1}$$

We use the entropy at a chosen decoding step t as our grounding objective, with t=1 corresponding to the first decoding step,

$$\mathcal{L}_{\text{ent}}(I,P) \;=\; H_t(I,P), \tag{2}$$

and backpropagate it to the projected visual embeddings $\mathbf{V} = \{\mathbf{v}_1,\ldots,\mathbf{v}_N\}$, i.e., the image tokens after the vision-language projector, where N denotes the number of image tokens. We now obtain the entropy-gradient signal as

$$\mathbf{G} \;=\; \frac{\partial \mathcal{L}_{\text{ent}}}{\partial \mathbf{V}}. \tag{3}$$

Intuitively, $\mathbf{G}$ measures how perturbations to each visual token would change the model's uncertainty. Since each token $\mathbf{v}_i$ corresponds to a spatial patch, we convert token-wise gradients into scalar saliency scores via the ℓ_2 norm,

$$s_i \;=\; \|\mathbf{G}_i\|_2\,, \tag{4}$$

yielding a score map $\mathbf{S} = \{s_1,\ldots,s_N\}$ that can be reshaped into an image-aligned heatmap. The resulting entropy-gradient map highlights regions whose visual evidence most affects the model's uncertainty about what to generate, serving as the primitive for multi-region extraction and the controlled refinement loop in the following sections.

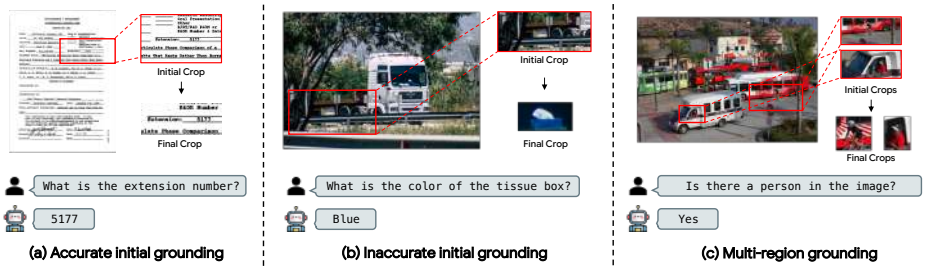

Fig. 5: Qualitative examples of iterative and multi-region grounding. (a) When the initial entropy-gradient crop already covers the evidence (e.g., a small document field), iterative refinement further zooms in to obtain a precise final crop and the correct answer. (b) Even when the initial crop is inaccurate, repeated re-grounding and refinement can recover and converge to the correct evidence region. (c) For queries requiring spatially disjoint cues, our method selects multiple regions of interest and refines them, enabling accurate prediction from multi-region evidence.

3.3 Multi-Region Selection and Ranking from Gradient Maps

The entropy-gradient score map $\mathbf{S}$ from Sec. 3.2 is defined on the visual-token grid and can contain small spurious peaks due to tokenization artifacts and backpropagation noise. Similar to [15], to obtain stable and spatially coherent regions of interest, we first smooth $\mathbf{S}$ with a Gaussian filter, producing a denoised map $\tilde{\mathbf{S}}$.

We then convert $\tilde{\mathbf{S}}$ into a binary support mask using an adaptive, data-driven threshold. Specifically, we sort the values of $\tilde{\mathbf{S}}$ and apply the elbow method [26] to select a threshold τ at the point of maximal deviation from the chord connecting the minimum and maximum values. This yields a parameter-free threshold that separates salient responses from background without requiring a manually tuned percentile. With the indicator function $\mathbb{I}[\cdot]$, the binary mask is computed as follows:

$$\mathbf{M}_i \;=\; \mathbb{I}[\tilde{s}_i \geq \tau]\,. \tag{5}$$

The resulting mask typically contains multiple disconnected components, each corresponding to a candidate region of interest. We extract connected components $\{C_j\}$ on the token grid and assign each component an importance score by accumulating the *original* saliency within the region:

$$w_j \;=\; \sum_{i \in C_j} s_i. \tag{6}$$

We rank components by w_j and retain the top-K regions. Each selected component is mapped back to the image as a tight bounding box, yielding a compact set of high-evidence, spatially coherent ROIs. These regions are then concatenated with the global view and used as input for final answer generation. The overall pipeline of acquiring binarized gradient maps is illustrated in Fig. 4.

3.4 Iterative Refinement

While entropy-based grounding can identify query-relevant regions, a single grounding pass is often insufficient in practice, as exemplified in Fig. 5. To address these issues, we introduce an iterative refinement procedure that repeatedly re-applies entropy-based grounding and region extraction on the current set of views. We initialize the view set with the original image, which is kept as a global context view, and the top-K regions from Sec. 3.3, ordered by their region scores. At each iteration, we run entropy-based grounding on each view, obtain a saliency map and a binarized mask, compute a tight bounding box around the activated support, and crop the corresponding patch. The cropped patches form the view set for the next iteration. This yields a simple test-time [9–11] feedback loop: grounding proposes where to look next, and re-invoking the model on these views enables either (i) deeper zoom-in for more decisive evidence or (ii) discovery of alternative regions that were overlooked in earlier iterations.

A key challenge is deciding when further refinement stops helping. We regulate the loop using *spatial entropy* [15]. Given a mask $\mathbf{M}$ with connected components $\{C_i\}_{i=1}^{N}$, we define

$$H(\mathbf{M}) = -\sum_{i=1}^{N} P(C_i) \log P(C_i), \tag{7}$$

where

$$P(C_i) = \frac{|C_i|}{\sum_{j=1}^{N} |C_j|}, \tag{8}$$

and $|C_i|$ denotes the number of active locations in component C_i. Spatial entropy measures dispersion: lower values indicate concentrated activation, while higher values indicate diffuse responses. In practice, we track the spatial entropy associated with the most important view. We continue refinement while this entropy decreases, and stop when it increases, indicating that further cropping no longer improves concentration and may start discarding useful context.

4 Experiments

4.1 Experimental Settings

Baselines. We evaluate our approach on four widely used VLM architectures: LLaVA-1.5 7B [20], LLaVA-1.6 Mistral 7B [21], InternVL-3.5 8B [31], and Qwen2.5-VL 7B [5]. For comparison, we include TEVA [13], a training-based grounding model, as well as SEAL [32]. We further evaluate the vanilla VLM backbones without our approach and compare against ViCrop [37] applied to LLaVA-1.5 and LLaVA-1.6. Additional evaluation details for these baselines are provided in the supplementary material.

Datasets. We evaluate our method on seven VQA benchmarks that stress *fine-grained and multi-region understanding*: TextVQA [29], V* [32], DocVQA [24],

Table 1: Quantitative results on standard reasoning benchmarks. Our training-free method improves fine-grained image understanding tasks across four VLM architectures. We include TEVA [13] and SEAL [32] as training-based references. Results for SEAL are taken from [37].

		TextVQA [29]	V* [32]	DocVQA [24]	POPE [19]	InfoQA [23]	GQA [12]	RWQA [34]
VLM	*Method*	*Fine-grained image understanding*					*General QA*	
Training-based								
TEVA	TEVA 3B [13]	66.80	61.60	-	87.00	-	-	-
	TEVA 7B [13]	72.50	77.10	51.3	87.90	31.90	-	-
Llava 1.5	SEAL [32]	36.30	75.30	5.31	82.40	-	50.18	-
Training-free								
LLaVA 1.5 [20]	Base model	46.22	46.07	22.32	86.55	22.24	**61.98**	**48.76**
	+ ViCrop [37]	**55.17**	47.64	19.63	87.25	**23.26**	60.97	47.97
	+ Ours	52.78	**56.02**	**33.70**	**87.56**	22.33	61.15	48.24
		+6.56	+9.95	+11.38	+1.01	+0.09	-0.76	-0.52
LLaVA 1.6 [21]	Base model	65.8	57.59	64.94	87.8	24.66	64.14	58.30
	+ ViCrop [37]	**68.65**	61.78	51.42	88.18	28.18	**64.54**	56.99
	+ Ours	67.96	**73.3**	**65.07**	**89.31**	**33.93**	63.97	**60.39**
		+2.16	+15.71	+0.13	+1.51	+9.27	-0.17	+2.09
InternVL 3.5 [31]	Base model	59.47	47.64	58.73	84.02	41.22	58.04	61.83
	+ Ours	**74.29**	**67.53**	**79.54**	**86.7**	**53.73**	**59.01**	**64.71**
		+14.82	+19.89	+20.81	+2.69	+12.51	+0.97	+2.88
Qwen 2.5 VL [5]	Base model	80.75	73.30	90.81	87.00	69.02	**61.01**	**67.84**
	+ Ours	**81.45**	**86.91**	**91.16**	**88.47**	**73.43**	59.49	66.93
		+0.70	+13.61	+0.35	+1.47	+4.41	-1.52	-0.91

POPE [19], InfoQA [23], and benchmarks that evaluate *general and real-world understanding*: GQA [12] and RWQA [34] to verify that our method does not harm general understanding capabilities. Together, they cover scene-text VQA, high-resolution visual search, document and infographic reasoning, compositional visual reasoning, hallucination probing, and real-world spatial understanding.

Evaluation Metrics. For TextVQA, we report the standard VQA accuracy metric.[1] For DocVQA and InfoQA, we submit predictions to the official evaluation server and report the official ANLS score returned by the test server. V*, GQA, and RWQA are evaluated using their standard top-1 accuracy. For POPE, we report accuracy averaged across all splits.

[1] `https://visualqa.org/evaluation.html`

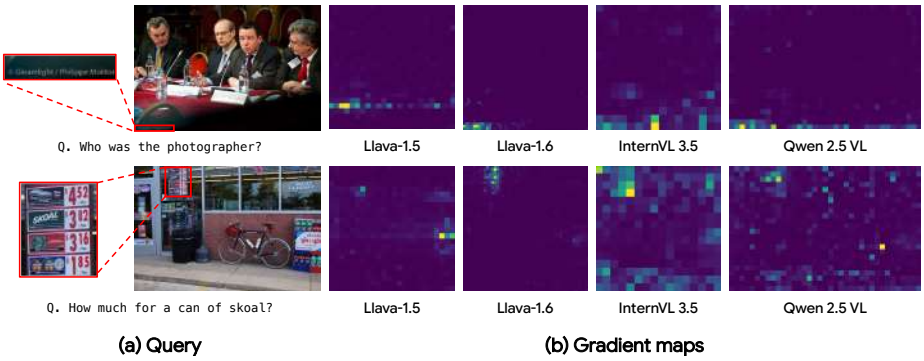

Fig. 6: Qualitative comparison of entropy-based gradient maps across different VLMs. We observe that weaker models (e.g., LLaVA 1.5) more frequently allocate gradients to spatially misaligned regions, whereas stronger models produce better-localized attributions. Among the compared models, LLaVA 1.6 generally yields the cleanest and most coherent gradient maps.

4.2 Experimental Results

Tab. 1 reports results on fine-grained image understanding [19, 23, 24, 29, 32] and general QA [12, 34]. We also show qualitative results in Fig. 5. In addition to downstream reasoning performance, we report quantitative localization metrics in the appendix.

Fine-grained understanding with consistent gains. Overall, our training-free grounding yields the most consistent gains on evidence-critical benchmarks where answers depend on small or spatially dispersed cues. We observe the largest gains on V* across all backbones, consistent with V* emphasizing visually crowded and high-resolution cases that benefit from targeted evidence acquisition. We also see considerable improvements in document and infoQA. Meanwhile, performance on general QA benchmarks, which primarily require global-level information, remains comparable, indicating that our method does not sacrifice broad scene understanding.

Comparison with baselines. Compared to ViCrop [37], which selects a single crop using heuristic scoring and a fixed crop ratio, our multi-region selection is considerably more robust on tasks requiring evidence aggregation from multiple locations. For example, ViCrop results in performance degradation on DocVQA, whereas applying our method brings substantial improvements. Furthermore, while TEVA [13] is a ***training-based*** method where its performance should be viewed as an upper bound, our training-free approach outperforms TEVA on tasks such as POPE and InfoQA, and reaches comparable results on V* owing to the significant improvements our grounding provides. We also observe that SEAL [32] is largely tailored to perform well on the V* totally failing in other datasets such as TextVQA or POPE.

Robustness across VLM architectures. Our method consistently improves performance across four VLM architectures, suggesting that entropy-gradient

Table 2: Ablation of different stopping criteria for iterative evidence retrieval.

Method	POPE	V*	DocVQA	RWQA	Inference Time
LLaVA 1.6 Base	87.8	57.59	64.94	58.30	**0.48 s**
Threshold-based					
Spatial Entropy	**89.31**	**73.3**	65.07	**60.39**	4.80 s
Confidence	89.07	70.16	64.26	58.69	3.84 s
Iteration-based					
1 Iteration	89.30	60.73	**65.42**	59.08	2.22 s
2 Iterations	89.21	65.96	63.34	**60.39**	2.70 s
3 Iterations	89.18	65.96	61.80	59.22	3.84 s

grounding generalizes beyond a specific vision encoder or language model. Fig. 6 further shows a qualitative trend: newer/stronger backbones tend to produce more accurately localized entropy-gradient maps, indicating that the quality of our model-intrinsic grounding signal scales with the underlying model's representations and uncertainty estimates.

4.3 Ablation Study

In this section, we present ablation studies and analyses of key design choices. Unless otherwise noted, all experiments are conducted with LLaVA-1.6 [21].

Stopping criterion and inference cost. We evaluate two families of stopping criteria for iterative evidence retrieval: *threshold-based* methods, which halt retrieval once a measured signal indicates sufficient evidence, and *iteration-based* methods, which run a fixed number of retrieval steps. For the threshold-based category, we compare our spatial-entropy criterion against a confidence-based alternative that stops when the maximum probability of the first generated token decreases, treating generation confidence as a proxy for evidence sufficiency. Tab. 2 summarizes the results. Among threshold-based methods, spatial entropy provides the best overall accuracy, achieving the highest scores on POPE and V* and matching the best RWQA result, while remaining competitive on DocVQA. The confidence-based heuristic is consistently weaker, suggesting that max-probability is susceptible to decoding instability and does not reliably indicate whether sufficient evidence has been gathered. Fixed iteration budgets reveal that no single iteration count is universally optimal: one iteration slightly improves DocVQA but underperforms on V*, while additional iterations can even degrade DocVQA. Although fixed budgets are faster than threshold-based stopping, they require manual tuning per dataset and still fail to match the best accuracy across benchmarks. Overall, spatial entropy offers a robust, adaptive stopping signal with a favorable accuracy–cost trade-off.

Effect of loss function for backpropagation. We ablate the objective used to generate gradient-based relevance maps in Tab.3. In addition to our default

Table 3: Ablation of the loss function used for gradient backpropagation.

Method	POPE	V*	DocVQA	RWQA
Entropy	**89.31**	**73.30**	65.07	**60.39**
Entropy Top-P	89.30	72.25	**65.09**	**60.39**
Maximum Probability	89.30	73.29	64.53	60.13

entropy objective, we consider two alternatives: (i) a top-P entropy variant that computes entropy over the minimal token set whose cumulative probability mass reaches 90%, and (ii) a maximum-probability objective that backpropagates the log of the top-1 probability. Overall, performance is stable across objectives, indicating that our localization signal is not overly sensitive to the exact choice of loss. Entropy achieves the best scores on POPE and V* and ties the best RWQA accuracy, while the top-p entropy slightly improves DocVQA with negligible changes elsewhere. In contrast, the maximum-probability objective is marginally worse on DocVQA and RWQA, suggesting that relying only on the top-1 token can discard useful distributional information when constructing relevance maps.

Table 4: Effects of varying the number of selected salient regions.

Method	POPE	V*	DocVQA	RWQA
LLaVA 1.6 Base	87.80	57.59	64.94	58.30
1 added	88.79	69.63	61.25	60.39
2 added	**89.31**	**73.30**	65.07	**60.39**
3 added	89.06	72.25	64.87	59.87
4 added	89.19	72.25	**65.14**	58.82

Effect of the number of selected regions. Tab. 4 ablates the number of appended regions K, i.e., how many top-ranked connected components are selected from the gradient map and added as additional views. Adding regions substantially improves fine-grained performance over the vanilla baseline, and we observe a clear optimum at K=2: performance peaks on POPE and V*, while also improving DocVQA and RWQA. While using a single region already boosts V* and RWQA, it can miss complementary evidence and even degrade DocVQA, consistent with document queries often requiring aggregation from multiple locations. Increasing K beyond two leads to only marginal and inconsistent changes in performance, suggesting diminishing returns as additional regions may introduce redundancy. Overall, selecting a small number of regions provides a good trade-off between capturing complementary evidence and avoiding redundant visual crops.

Layer-wise gradient analysis. We vary the transformer layer used to compute the entropy-gradient signal and report the performance in Tab. 5. It shows that shallow layers perform poorly, while deeper layers yield markedly stronger

Table 5: Effects of different layers for gradient computation.

Method	POPE	V*	DocVQA	RWQA
Layer 15	88.32	58.64	50.99	58.17
Layer 20	88.84	**78.01**	62.68	**61.18**
Layer 26	89.07	72.25	64.75	60.78
Layer 32 (Last)	**89.31**	73.30	**65.07**	60.39

Table 6: Performance when backpropagating the loss from different generated tokens. Results are reported on [24, 29], as these datasets require multi-token generation.

Dataset	First	Second	Third	Fourth
TextVQA	**67.96**	65.27	67.37	66.35
DocVQA	65.07	63.70	**65.28**	64.64

results. The final layer is the most robust overall, consistent with Fig. 7 where gradient magnitudes increase toward deeper layers. We therefore use the last layer in all experiments; earlier layers are omitted due to consistently weak performance.

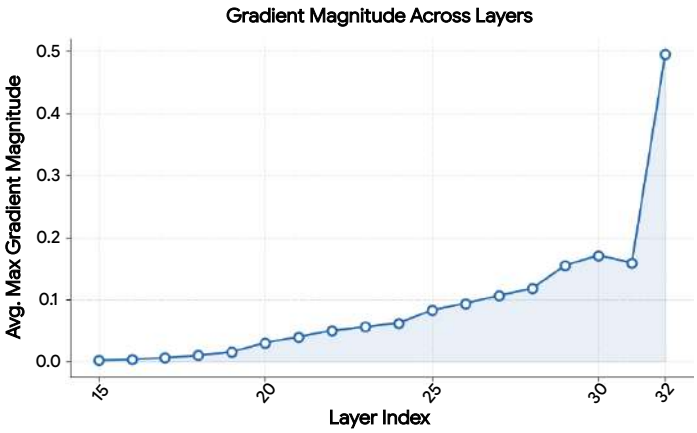

Fig. 7: Layer-wise maximum average gradient magnitude on TextVQA, illustrating how gradient strength evolves with model depth.

Subsequent-token loss backpropagation. We study whether gradients from later decoding steps provide a better grounding signal than those from the first step. Tab. 6 reports results when backpropagating the loss from the t-th generated token ($t \in \{1, 2, 3, 4\}$) while keeping all other components fixed. Backpropagating from the first token performs best overall: using later tokens generally degrades TextVQA and does not yield consistent improvements on DocVQA (the third token provides a slight gain, but the trend is not stable across to-

kens). As visualized in Fig. 8, gradients from subsequent tokens become increasingly conditioned on previously generated tokens, which is expected under causal self-attention and can bias the relevance map toward an already committed interpretation. In addition, using token $t > 1$ requires unrolling multiple decoding steps prior to backpropagation, increasing compute and memory. We therefore use the first token as a simple and efficient choice in all experiments.

Fig. 8: Example of gradient computation at different token positions. Gradients taken at later tokens show stronger conditioning on specific regions, which can lead to better localization but also reduces exploration of other areas.

Computational overhead. We report the additional inference time introduced by our method as the average per-sample runtime on RealWorldQA [34] (726–1536 px images), shown in Tab. 7. We report both a single retrieval pass and the full iterative variant with our stopping criterion. As expected, iterative evidence retrieval increases runtime due to repeated forward/backward passes. Despite this overhead, runtimes remain practical, and the stopping criterion limits unnecessary iterations, yielding a bounded cost while providing the accuracy gains reported in Tab. 1.

Table 7: Inference time comparison between a single and iterative runs.

Method	LLaVA-1.5	LLaVA-1.6	InternVL-3.5	QWENVL-2.5
Single iteration	1.00 s	2.04 s	1.18 s	1.18 s
Iterative (w/ stopping)	3.10 s	3.37 s	3.17 s	2.98 s

5 Conclusion

In this work, we introduced a training-free, model-intrinsic visual grounding framework for pretrained VLMs by backpropagating the entropy of the next-token distribution to visual embeddings. Using uncertainty gradients as a decision-relevant signal and converting them into ranked regions of interest, our method retrieves evidence from spatially disjoint cues without auxiliary detectors or heuristic attention processing. To address fixed-resolution limitations, we further propose an iterative refinement loop guided by a spatial-entropy stopping

criterion, enabling the model to acquire finer-grained evidence and recover overlooked regions at inference time. Extensive experiments across standard reasoning benchmarks and four VLM architectures show consistent improvements on evidence-critical tasks—particularly in high-resolution and document-centric settings—while producing more focused, query-conditioned localizations.

References

1. An, H., Jung, J., Kim, M., Hong, S., Kim, C., Fukuda, K., Jeon, M., Han, J., Narihira, T., Ko, H., et al.: C3g: Learning compact 3d representations with 2k gaussians. arXiv preprint arXiv:2512.04021 (2025)
2. An, H., Kim, J.H., Park, S., Jung, J., Han, J., Hong, S., Kim, S.: Cross-view completion models are zero-shot correspondence estimators. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1103–1115 (2025)
3. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., Wu, C., Tan, H., Li, C., Yang, J., Yu, J., Wang, X., Qin, B., Wang, Y., Yan, Z., Feng, Z., Liu, Z., Li, B., Deng, J.: Llava-onevision-1.5: Fully open framework for democratized multimodal training (2025), https://arxiv.org/abs/2509.23661
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
6. Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., Lu, J., Anderson, T., Bransom, E., Ehsani, K., Ngo, H., Chen, Y., Patel, A., Yatskar, M., Callison-Burch, C., Head, A., Hendrix, R., Bastani, F., VanderBilt, E., Lambert, N., Chou, Y., Chheda, A., Sparks, J., Skjonsberg, S., Schmitz, M., Sarnat, A., Bischoff, B., Walsh, P., Newell, C., Wolters, P., Gupta, T., Zeng, K.H., Borchardt, J., Groeneveld, D., Dumas, J., Nam, C., Lebrecht, S., Wittlif, C., Schoenick, C., Michel, O., Krishna, R., Weihs, L., Smith, N.A., Hajishirzi, H., Girshick, R., Farhadi, A., Kembhavi, A.: Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. arXiv preprint arXiv:2409.17146 (2024)
7. Gurbuz, A.S., Hong, S., Nassar, A., Pollefeys, M., Staar, P.: Moving beyond sparse grounding with complete screen parsing supervision. arXiv preprint arXiv:2602.14276 (2026)
8. Han, J., Hong, S., Jung, J., Jang, W., An, H., Wang, Q., Kim, S., Feng, C.: Emergent outlier view rejection in visual geometry grounded transformers. arXiv preprint arXiv:2512.04012 (2025)
9. Hong, S., Cho, S., Kim, S., Lin, S.: Unifying feature and cost aggregation with transformers for semantic and visual correspondence. arXiv preprint arXiv:2403.11120 (2024)
10. Hong, S., Kim, S.: Deep matching prior: Test-time optimization for dense correspondence. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9907–9917 (2021)

11. Hong, S., Nam, J., Cho, S., Hong, S., Jeon, S., Min, D., Kim, S.: Neural matching fields: Implicit representation of matching fields for visual correspondence. Advances in Neural Information Processing Systems **35**, 13512–13526 (2022)
12. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6693–6702 (2019). https://doi.org/10.1109/CVPR.2019.00686
13. Jiang, Y., Gu, J., Xue, T., Cheung, K.C., Molchanov, P., Yin, H., Liu, S.: Token-efficient vlm: High-resolution image understanding via dynamic region proposal. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 24147–24158 (October 2025)
14. Kaduri, O., Bagon, S., Dekel, T.: What's in the image? a deep-dive into the vision of vision language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14549–14558 (2025)
15. Kang, S., Kim, J., Kim, J., Hwang, S.J.: Your large vision-language model only needs a few attention heads for visual grounding (2025), https://arxiv.org/abs/2503.06287
16. Kim, C., Shin, H., Hong, E., Yoon, H., Arnab, A., Seo, P.H., Hong, S., Kim, S.: Seg4diff: Unveiling open-vocabulary segmentation in text-to-image diffusion transformers. arXiv preprint arXiv:2509.18096 (2025)
17. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9579–9589 (June 2024)
18. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. Transactions on Machine Learning Research (2024)
19. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023), https://openreview.net/forum?id=xozJw0kZXF
20. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2024), https://arxiv.org/abs/2310.03744
21. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
22. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection (2024), https://arxiv.org/abs/2303.05499
23. Mathew, M., Bagal, V., Tito, R.P., Karatzas, D., Valveny, E., Jawahar, C.V.: Infographicvqa (2021), https://arxiv.org/abs/2104.12756
24. Mathew, M., Karatzas, D., Jawahar, C.V.: Docvqa: A dataset for vqa on document images (2021), https://arxiv.org/abs/2007.00398
25. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection (2016), https://arxiv.org/abs/1506.02640
26. Satopaa, V., Albrecht, J., Irwin, D., Raghavan, B.: Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In: 2011 31st international conference on distributed computing systems workshops. pp. 166–171. IEEE (2011)
27. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)

28. Shen, H., Zhao, K., Zhao, T., Xu, R., Zhang, Z., Zhu, M., Yin, J.: ZoomEye: Enhancing multimodal LLMs with human-like zooming capabilities through tree-based image exploration. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 6602–6618. Association for Computational Linguistics, Suzhou, China (Nov 2025). https://doi.org/10.18653/v1/2025.emnlp-main.335, https://aclanthology.org/2025.emnlp-main.335/
29. Singh, A., Natarjan, V., Shah, M., Jiang, Y., Chen, X., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8317–8326 (2019)
30. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328. PMLR (06–11 Aug 2017), https://proceedings.mlr.press/v70/sundararajan17a.html
31. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
32. Wu, P., Xie, S.: V*: Guided visual search as a core mechanism in multimodal llms (2023), https://arxiv.org/abs/2312.14135
33. Wu, S., Jin, S., Zhang, W., Xu, L., Liu, W., Li, W., Loy, C.C.: F-lmm: Grounding frozen large multimodal models. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24710–24721 (2025). https://doi.org/10.1109/CVPR52734.2025.02301
34. xAI: Realworldqa dataset (2024), https://huggingface.co/datasets/xai-org/RealworldQA, dataset card (accessed 2026-02-28)
35. Xie, Y., Yang, K., An, X., Wu, K., Zhao, Y., Deng, W., Ran, Z., Wang, Y., Feng, Z., Miles, R., Elezi, I., Deng, J.: Region-based cluster discrimination for visual representation learning. In: ICCV (2025)
36. Yoon, H., Jung, J., Kim, J., Choi, H., Shin, H., Lim, S., An, H., Kim, C., Han, J., Kim, D., et al.: Visual representation alignment for multimodal large language models. arXiv preprint arXiv:2509.07979 (2025)
37. Zhang, J., Khayatkhoei, M., Chhikara, P., Ilievski, F.: Mllms know where to look: Training-free perception of small visual details with multimodal llms (2025), https://arxiv.org/abs/2502.17422
38. Zhang, Z., Yadav, S., Han, F., Shutova, E.: Cross-modal information flow in multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19781–19791 (2025)

Entropy-Gradient Grounding: Training-Free Evidence Retrieval in Vision-Language Models

–Appendix–

Table of Contents

A Pytorch-Style Psuedo Code

We provide a Pytorch-Style psuedo code in Alg. 1.

Algorithm 1: Entropy-Driven Multi-Region Localization with Iterative Refinement

Input: Image I, prompt P, VLM $\mathcal{M}$, max iterations T, top regions K
Output: Refined image set $\mathcal{I}$
Initialize $\mathcal{I}_0 \leftarrow \{I\}$
Initialize $H_{\text{prev}} \leftarrow +\infty$
for $t = 0$ **to** $T - 1$ **do**
  Initialize candidate component set $\mathcal{C} \leftarrow \emptyset$
  foreach $I_i \in \mathcal{I}_t$ **do**
    Compute next-token distribution $p(y \mid I_i, P)$
    Compute entropy loss $\mathcal{L}_{\text{ent}}$
    Backpropagate $\mathcal{L}_{\text{ent}}$ to visual embeddings
    Obtain saliency map $\mathbf{S}_i$
    Extract connected components $\{C_{i,j}\}$ from $\mathbf{S}_i$
    foreach $C_{i,j}$ **do**
      Compute component score $w_{i,j}$
      Add $(I_i, C_{i,j}, w_{i,j})$ to $\mathcal{C}$
  Select top-K components from $\mathcal{C}$ by score
  Let $(I^\star, C^\star)$ be the highest-scoring component
  Compute spatial entropy H_t of $C^\star$
  if $H_t \geq H_{prev}$ **then**
    break
  $H_{\text{prev}} \leftarrow H_t$
  Initialize $\mathcal{I}_{t+1} \leftarrow \{I\}$ `// retain global context`
  foreach *selected* (I_i, C_i) **do**
    Crop bounding box of C_i from I_i to obtain I_{C_i}
    Add I_{C_i} to $\mathcal{I}_{t+1}$
return $\mathcal{I}_t$

B Evaluation Details

For each baseline model [5, 20, 21, 31], we retain its default inference configuration and apply our method on top without modifying the baseline itself. For example, LLaVA-1.5 takes a single input image by default, whereas LLaVA-1.6 uses five images. Since our approach naturally accommodates varying numbers of input images, we preserve these settings as is. Similarly, InternVL and Qwen use original-resolution images in their default inference setups, and our method remains fully compatible with this setting, allowing us to integrate it directly without any additional changes.

C Dataset Information

This section briefly summarizes the datasets used in our evaluation.

- *TextVQA* [29]: A scene-text VQA dataset where answering questions requires reading and reasoning about text embedded in natural images, making it particularly sensitive to small, detail-critical visual evidence.
- *V** [32]: A benchmark introduced alongside the V* guided visual search framework to evaluate detail-focused understanding in visually crowded and high-resolution scenes, where relevant cues are often small and easily overlooked.
- *DocVQA* [24]: A document-image VQA dataset requiring text- and layout-aware reasoning across diverse document types (e.g., forms, receipts, and papers), often involving locating multiple pieces of information within a page.
- *InfoQA* [23]: A VQA benchmark on infographic images requiring joint reasoning over textual content, layout, graphics, and data visualizations, frequently involving elementary arithmetic and multi-step evidence aggregation.
- *GQA* [12]: A large-scale compositional VQA benchmark designed for structured visual reasoning, where questions depend on relationships between multiple objects and attributes.
- *POPE* [19]: A polling-based probing benchmark designed to evaluate *object hallucination* in VLMs using controlled yes/no queries about object presence.
- *RWQA* [34]: A real-world reasoning benchmark introduced with Grok-1.5V, consisting of images from everyday and vehicle-captured scenarios with verifiable answers, emphasizing spatial and physical reasoning in real-world environments.

Furthermore, TextVQA, InfoQA, and DocVQA provide additional OCR tokens extracted by a third-party model to support inference. In our evaluations, we do not utilize these tokens.

D Comparison with Additional Methods

Here, we further compare against ZoomEye [28] on LLaVA-1.5 7B, the only backbone shared by both works. Although both approaches are also compatible with Qwen 2.5, ZoomEye is evaluated with the 3B variant, whereas our method uses the 7B variant. In our experiments, running ZoomEye on Qwen 2.5 7B was infeasible due to out-of-memory issues, even on an A100 GPU with 80GB VRAM.

ZoomEye is a training-free, model-agnostic tree-search method that represents an image as a hierarchical tree, where the root denotes the full image and each child node corresponds to a zoomed-in sub-region of its parent. Given a question, it guides the MLLM to traverse this tree by assigning confidence-based priority scores to candidate nodes, effectively mimicking human-like zoom-in behavior to identify task-relevant visual evidence. The search terminates once the model's answer confidence exceeds a predefined threshold.

Table 8: Quantiative comparison to ZoomEye on Llava 1.5.

	TextVQA	V*	DocVQA	POPE	InfoQA	GQA	RWQA
Base model	46.22	46.07	22.32	86.55	22.24	61.98	48.76
ZoomEye	46.68	72.25	24.61	86.85	23.23	61.42	49.67
Ours	52.78	56.02	33.70	87.56	22.33	61.15	48.24

We compare ours with ZoomEye across several benchmarks in Tab. 8. While ZoomEye achieves strong performance on the V* benchmark, which consists of high-resolution images with fine-grained visual elements that benefit from its exhaustive tree-based exploration, it does not generalize as effectively to other datasets. In contrast, our method demonstrates more consistent improvements across diverse reasoning tasks, particularly on TextVQA and DocVQA. This pattern is similar to what we observe with SEAL [32]: methods that are specifically designed for high-resolution visual search scenarios tend to excel on benchmarks like V* but may struggle on tasks that require different forms of visual reasoning, such as document understanding or scene-text recognition. Our entropy-gradient grounding, by contrast, provides a more general-purpose grounding signal that adapts to a wider range of query types and visual contexts without relying on task-specific search heuristics.

E Quantitative Evaluation of Localization Accuracy

While the main paper demonstrates qualitative grounding results and quantitative improvements on downstream tasks, it is equally important to directly assess the localization accuracy of our grounding signal. To this end, we measure the Intersection-over-Union (IoU) between the spatial regions identified by our entropy-gradient mask and the ground-truth bounding boxes for TextVQA annotated by ViCrop [37]. All evaluations are conducted with LLaVA 1.6.

Tab. 9 shows that our method achieves substantially higher IoU than ViCrop, indicating more accurate localization of query-relevant visual evidence. This result is consistent with the qualitative comparisons in Fig. 3 and Fig. 5, and provides direct quantitative evidence that backpropagating entropy to visual embeddings yields a grounding signal better aligned with the spatial evidence necessary for answering the question.

Table 9: Quantitative localization performance on TextVQA using ground-truth bounding boxes provided by ViCrop.

Metric	Ours	ViCrop
IoU	0.29	0.14

F Qualitative Comparison with Additional Gradient-based Grounding Methods

While our method backpropagates the entropy of the next-token distribution to the projected visual embeddings $\mathbf{V}$, alternative choices exist for both the backpropagation target and the objective. For example, one could backpropagate gradients to raw image pixels instead of visual embeddings, or optimize the log-probability of the top-1 token rather than Shannon entropy.

ViCrop [37] considers such an alternative in its `pure-grad` ablation. For each image–question pair (x, q), it computes the log of the maximum output probability at the first answer token and visualizes the resulting gradients after taking the ℓ_2 norm across image channels. Our method differs in two key ways. First, we backpropagate to *visual embeddings* rather than raw pixels, which yields semantically richer gradients that are better aligned with the model's internal representation space. Second, we optimize the Shannon entropy of the full next-token distribution, which captures the model's global uncertainty over the vocabulary, rather than the log-probability of a single predicted token, which ignores the rest of the distribution.

Fig. 9 shows that `pure-grad` produces diffuse and poorly structured gradient maps, making query-relevant regions difficult to identify. By contrast, our entropy-gradient maps are spatially compact and well localized, clearly highlighting the visual evidence needed to answer the question. These qualitative comparisons support our design choice of backpropagating entropy to visual embeddings as a more effective grounding strategy.

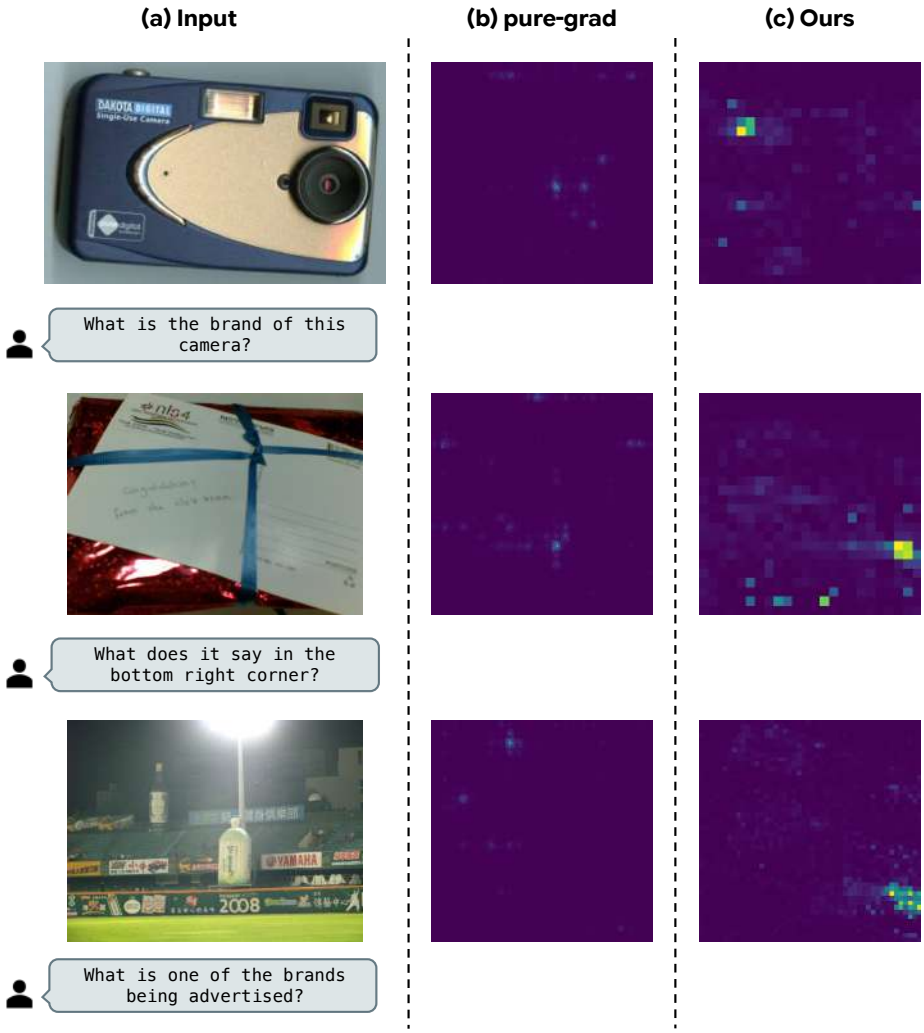

Fig. 9: Qualitative comparison of gradient-based grounding methods on LLaVA 1.5. For each example, we show (a) the input image with the corresponding question, (b) the gradient map produced by the `pure-grad` method of ViCrop [37], which backpropagates the log-probability of the top-1 token to the input pixels, and (c) the entropy-gradient map produced by our method, which backpropagates the Shannon entropy to the visual embeddings. The `pure-grad` maps appear diffuse and lack discernible spatial structure, whereas our entropy-gradient maps produce well-localized activations that concentrate on the query-relevant regions, confirming the effectiveness of operating in the embedding space with an entropy-based objective.

G Prompts

In this section, we list the prompts used during inference for all evaluated models. We report the exact phrasing used for the questions, while the full prompt structure is generated using each model's native prompt template. Among the datasets we use for evaluation, TextVQA, DocVQA, InfoVQA, GQA, and POPE datasets are open-ended questions, where V^* and RWQA datasets ask the model to answer from multiple choices. For each category, we use the prompts formatted as follows:

– **Open-ended questions**

Prompt Template

```
{Question}
Answer the question using a single word or phrase.
```

– **Multiple-choice questions**

Prompt Template

```
{Question}
(A) ...
(B) ...
(C) ...
(D) ...
Answer with the option's letter from the given choices
directly.
```

H Additional Qualitative Examples

In this section, we provide additional qualitative examples of the entropy-gradient map and the final crop after our interactive refinement applied to various baselines. The examples are shown in Fig. 10, Fig. 11, and Fig. 12.

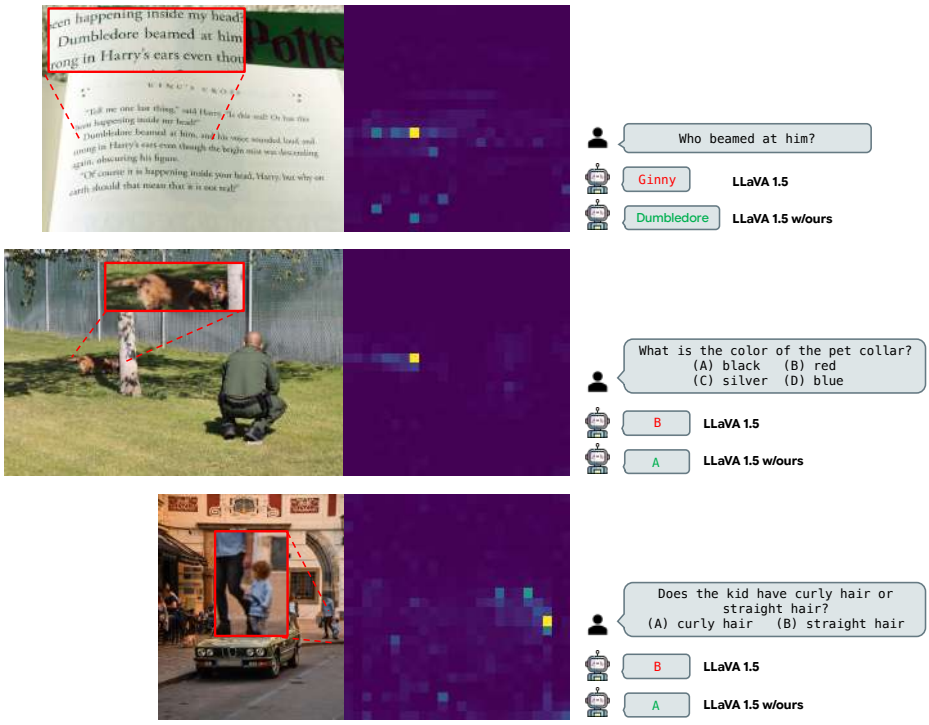

Fig. 10: Additional Examples. Qualitative examples on Llava 1.5. The most important final crop is highlighted in red.

Fig. 11: Additional Examples. Qualitative examples on Llava 1.6. The most important final crop is highlighted in red.

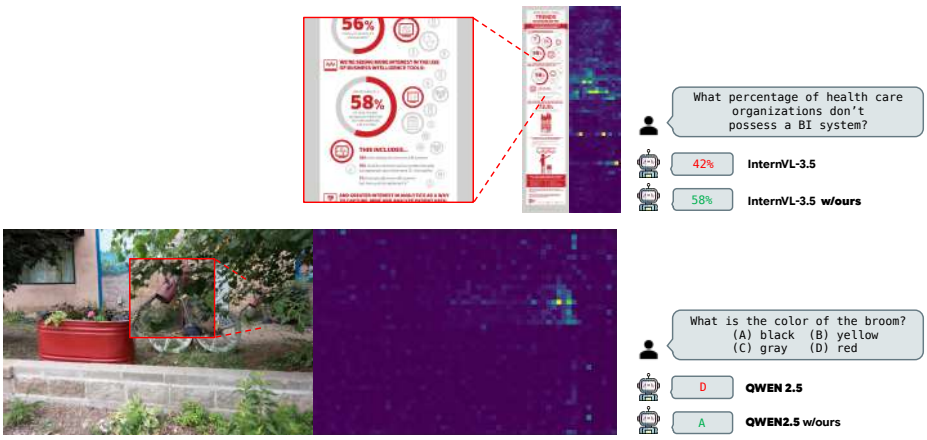

Fig. 12: Additional Examples. Qualitative examples on native resolution encodings (QWEN 2.5, InternVL 3.5). The most important final crop is highlighted in red.

I Limitations

Although our method improves the localization of query-relevant visual evidence, correct localization by itself is not sufficient to ensure a correct final answer. The downstream language model must still accurately interpret the visual content and carry out the required reasoning, and thus errors may persist even when the relevant region is localized correctly. As shown in Fig. 13, the model can still misinterpret a detailed crop that zooms into the correct location to answer the question. Furthermore, our method may fail to recover relevant regions when the backbone model itself does not consider them important, since it ultimately relies on the spatial signals produced by the underlying model (see Fig. 6).

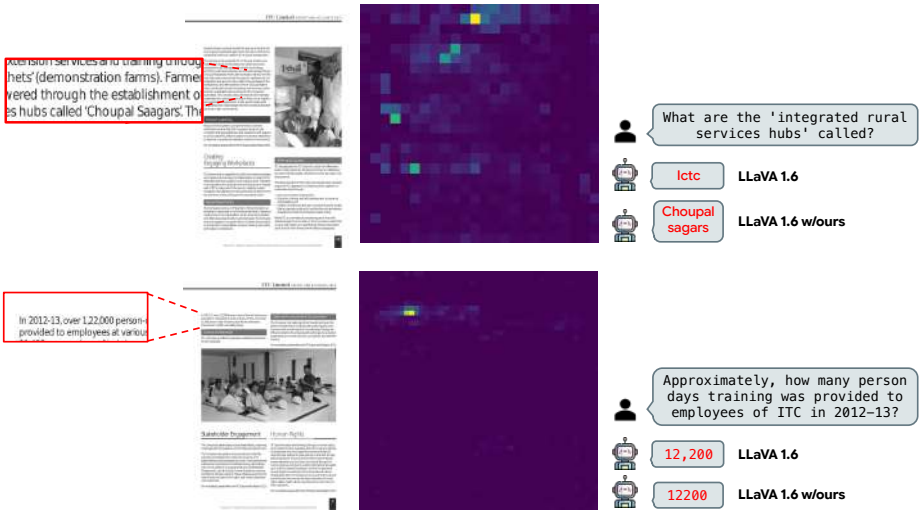

Fig. 13: Failure cases. Qualitative examples of failing to predict the correct answer illustrating a limitation of our method. Even when confronted with a very detailed crop, the model still fails to answer correctly.